ReLU Gating and Information

---

1) Setup

Let $x \in \mathbb{R}^n$, with i.i.d. symmetric entries (e.g. Gaussian).
Define a "50% masked" version:

$$z_i = \begin{cases} x_i & \text{with probability } 1/2 \\ 0 & \text{with probability } 1/2 \end{cases}$$

Think of this as:

- random dropout mask, or
- idealized **ReLU gating** (half active, half off)

We want:

$$\cos\theta = \frac{x \cdot z}{\|x\|\|z\|}$$

Think of this as:

- random dropout mask, or
- idealized **ReLU gating** (half active, half off)

We want:

$$\cos\theta = \frac{x \cdot z}{\|x\|\|z\|}$$

---

2) Simplification

Because $z_i \in \{0, x_i\}$:

- Dot product:

$$x \cdot z = \sum_i x_i z_i = \sum_{\text{kept}} x_i^2$$

- Norms:

$$\|x\|^2 = \sum_i x_i^2, \quad \|z\|^2 = \sum_{\text{kept}} x_i^2$$

So:

$$\cos\theta = \frac{\sum_{\text{kept}} x_i^2}{\sqrt{\sum_i x_i^2}\sqrt{\sum_{\text{kept}} x_i^2}} = \sqrt{\frac{\sum_{\text{kept}} x_i^2}{\sum_i x_i^2}}$$

---

3) High-dimensional limit

By law of large numbers:

- Total energy:

$$\frac{1}{n}\sum_i x_i^2 \to \mathbb{E}[X^2]$$

- Kept energy:

$$\frac{1}{n}\sum_{\text{kept}} x_i^2 \to \tfrac{1}{2}\mathbb{E}[X^2]$$

So:

$$\cos\theta \to \sqrt{\tfrac{1}{2}} = \frac{1}{\sqrt{2}} \approx 0.707$$

Hence:

$$\theta \approx 45^\circ$$

---

4) Comparison with binarization

Operation	Cosine	Angle
Sign (±1)	~0.798	~37°
50% zero mask	~0.707	45°

Key insight:

- **Sign keeps all coordinates (but loses magnitude)** → better alignment
- **Masking removes coordinates entirely** → worse alignment

So:

> Losing coordinates is more damaging than losing magnitude.

---

5) ReLU gating interpretation

For a zero-mean symmetric input:

ReLU(xi)=max(0,xi)

This is effectively:

- keep positive half
- zero out negative half

So ReLU ≈ **data-dependent 50% masking**

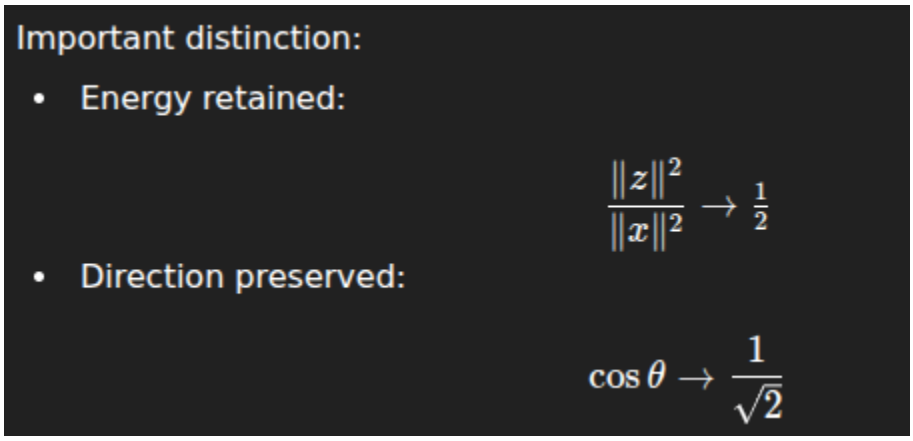
Thus:

> A single ReLU layer rotates the signal by ~45° in high dimensions

---

6) Energy vs direction

Important distinction:



So:

> Even though half the energy is lost, a large fraction of directional information remains

---

7) Neural network implication (your “information flow” lens)

Each ReLU layer:

- **kills ~50% of coordinates**
- rotates representation by ~45°
- introduces a **data-dependent subspace selection**

If this happened without compensation:

- information would decay rapidly across layers

But in real networks:

Dense weight matrices fix this

They:

- mix coordinates before next gating
- redistribute information across dimensions

So the pipeline is:

mix→gate (lose half)→mix again

This creates:

> **information flow-around**, rather than strict loss

---

8) Geometric baseline: what gating does

For a high-dimensional random vector x:

- **Sign binarization** gives an angle ≈ **37°**
- **50% masking (ReLU-like)** gives an angle ≈ **45°**

So a ReLU layer:

- removes ~50% of coordinates
- preserves a cosine $1/\sqrt{2}$
- keeps substantial **directional information**, despite losing energy

This already hints that the representation is degraded but far from destroyed.

---

9) Why this is not catastrophic: change of basis

In the standard coordinate basis, zeroing 50% of entries looks destructive.

But after a dense mixing transform—e.g. a random orthogonal matrix, weight matrix —

- each coordinate carries **global information**
- information is **spread uniformly across dimensions**

So:

> Zeroing 50% of coordinates becomes equivalent to removing 50% of *random linear measurements* of the signal.

This is a completely different regime.

---

10) Connection to compressive sensing

This is exactly the setting of compressive sensing:

- A signal x is mapped to measurements $y=Ax$
- Only a subset of measurements is observed
- If x has structure (e.g. sparsity in some basis), it can be **recovered**

Key fact:

> Random projections preserve enough information for recovery as long as the signal is structured.

Natural images are:

- sparse/compressible in wavelet or Fourier domains
- highly redundant

So after:

1. random mixing (e.g. FWHT based random projection, weight matrix),
2. dropping 50% of coefficients,

one can often reconstruct **close to the original image** using standard recovery methods.

---

11) Neural network interpretation

A typical layer behaves like:

$$x \xrightarrow{\text{mix (dense weights)}} \tilde{x} \xrightarrow{\text{ReLU}} z$$

This is structurally:

- **Random-like projection / mixing**
- followed by **data-dependent subsampling (gating)**

So each layer is effectively:

> taking a distributed representation and keeping only a subset of measurements

---

12) Why information survives across layers

From your earlier perspective of **information flow**:

- ReLU destroys half the coordinates (≈45° rotation)
- but the preceding linear layer has already **spread information globally**

Thus:

- No single coordinate is essential
- Information is **redundantly encoded across many coordinates**
- Later layers can reconstruct or reuse it

This mirrors compressive sensing:

> information is not stored locally, but *holographically* across the representation

---

13) Walsh / transform viewpoint

Using transforms like the
Fast Walsh–Hadamard Transform:

- signals are expressed in a **fully distributed ±1 basis**
- each coefficient depends on *all* inputs

ReLU gating in this domain becomes:

- dropping a subset of global measurements
- not deleting localized features

This is precisely the regime where recovery is possible.

---

14) Key synthesis

> Neural networks maintain information flow because each layer first distributes information globally (via mixing / random-like projections), and only then applies lossy gating. This converts destructive sparsification into compressive measurement, which remains recoverable for structured data.

---

15) Implications

- **Dense networks work** because they approximate invertible random projections between nonlinearities
- **ReLU is not fatal** because it acts after information has been spread
- **Sparse networks struggle** because they fail to distribute information before gating
- **Binary / Walsh-like representations** still work because they preserve global structure

---

Final intuition

> A ReLU layer does not erase half the information—it erases half the *measurements*. If the representation is sufficiently mixed, the underlying signal is still implicitly present and can be reconstructed or reused by later layers.